Classifier Fusion for SVM-Based Multimedia Semantic Indexing

Stéphane Ayache, Georges Quénot, Jérôme Gensel

Laboratoire d'Informatique de Grenoble (LIG)
385 rue de la Bibliothque - BP 53
38041 Grenoble - Cedex 9

Abstract. Concept indexing in multimedia libraries is very useful for users searching and browsing but it is a very challenging research problem as well. Combining several modalities, features or concepts is one of the key issues for bridging the gap between signal and semantics. In this paper, we present three fusion schemes inspired from the classical early and late fusion schemes. First, we present a kernel-based fusion scheme which takes advantage of the kernel basis of classifiers such as SVMs. Second, we integrate a new normalization process into the early fusion scheme. Third, we present a contextual late fusion scheme to merge classification scores of several concepts. We conducted experiments in the framework of the official TRECVID'06 evaluation campaign and we obtained significant improvements with the proposed fusion schemes relatively to usual fusion schemes.

1 Introduction

In order to retrieve multimedia documents from huge digital libraries, the needs for concept-based indexing are rapidly growing. Finding concepts in multimedia documents, such as video sequences, is one of the main objectives of the content-based semantic indexing community. Hence, new issues are arising on the combination (fusion) of several features, modalities and/or intermediate concepts to obtain a better accuracy of concept detection. For instance, an efficient fusion scheme must enhance concept indexing in multimedia documents by merging visual and textual modalities, color and texture modalities, or global and local features. Using a generic framework, usual approaches propose either to merge data on a concatenated vector before achieving classification [1, 2], or to perform several classification and then to merge confidence scores using a higher level classifier [6, 11] by the means of a stacking technique [16]. Called "early" and "late" fusion [13], those approaches are easy to implement and provide state of the art performance. However, such fusion schemes are not always able to outperform unimodal classifiers, especially when one of the modalities provide much better accuracy than the others or when one has to handle imbalanced input features. Such situations are particularly frequent in the field of multimedia indexing due to the diversity of concepts with regard to the extracted features.

Using kernel-based classifier, for instance a Support Vector Machine, recent approaches have been proposed to take advantage of some useful kernel properties. They aim to merge features at the kernel level before performing the concept classification. Kernel-based data fusion has been successfully applied in biology to the problem of predicting the function of yeast proteins [8]. [8, 15] propose efficient algorithms to learn simultaneously the parameters of each unimodal kernel and the parameters of the combining function.

In this paper, we study and compare three fusion schemes in the scope of semantic video indexing. The first one takes advantage of some useful kernel properties, we present a simple algorithm which merges unimodal kernels before performing the concept classification using a SVM classifier. In such a way, features are combined at the earliest possible step using a kernel-based classifier. The second fusion scheme is derived from the early fusion scheme. We normalized each individual feature vectors so that their average norm becomes equal in order to reduce the problem of imbalanced input features. The third fusion scheme is a late-like fusion scheme; it performs fusion at the concept level taking into account the classification scores of 39 concepts from visual and textual modalities.

In Section 2, we briefly present Support Vectors Machines and some required knowledge about kernels. In section 3, we describe the proposed fusion schemes and give some background information for formally comparing them with other fusion schemes. In section 4, we describe the experiments conducted using the TRECVID'06 [7] corpus and metrics.

2 Kernel-based classifier

Kernel-based methods have provided successful tools for solving many recognition problems, such as KFD, KPCA or SVM [12]. One of the reasons of this success is the use of kernels which overcome the problem of non-linearly separable data sets by mapping the initial problem into a higher dimensional space. The main idea behind kernel-based classifiers is that the similarity between examples in a data set gives much information about the patterns that may be present in these data.

2.1 Support Vector Machines

Support Vector Machines (SVM) have shown their capacities in pattern recognition and have been widely used for classification in CBIR. SVM is formalized as an optimization problem which finds the best hyperplane separating relevant and irrelevant vectors by maximizing the size of the margin between both sets. The use of a kernel allows the algorithm to find the maximum-margin hyperplane in a transformed feature space. The transformation may be non-linear and the transformed space may be of higher dimensionality than the original one. Thus, though the classifier separator is a hyperplane in the high-dimensional feature space it may be non-linear in the original input space. Furthermore, if the kernel

used is a Gaussian radial basis function, the corresponding feature space is a Hilbert space of infinite dimension. Maximum margin classifiers are well regularized and the infinite dimension does not spoil the results. In a two-class case, the decision function for a test sample $\mathbf{x}$ has the following form:

$$g(\mathbf{x}) = \sum_i \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) - b$$

where $K(\mathbf{x}_i, \mathbf{x})$ is the value of a kernel function for the training sample $\mathbf{x}_i$ and the test sample $\mathbf{x}$, y_i the class label of $\mathbf{x}_i$ ($+1$ or -1), α_i the learned weight of the training sample $\mathbf{x}_i$, and b is a learned threshold parameter. The training samples with weight $\alpha_i > 0$ are usually called "support vectors".

2.2 Kernel matrices

A kernel matrix is a similarity matrix where each entry represents a measure of similarity between two sample vectors, and must be positive definite (i.e. satisfy the Mercer's condition) to ensure that the optimization problem is convex. Therefore, a clever choice of the kernel (or similarity) function is essential to obtain a positive definite kernel matrix. Nevertheless, some unproved positive definite kernels such as EMD-based kernels or Log kernels have been successfully used in image recognition [18, 3].

Kernel matrices satisfying the Mercer's condition have interesting properties which provide modularity and derivation of kernels from other kernels. If K and K' are kernels, then the following are also kernels (not exhaustive):

- $aK + bK'$, for $a > 0$ and $b \geq 0$
- $K \times K'$, where $\times$ is the entrywise product

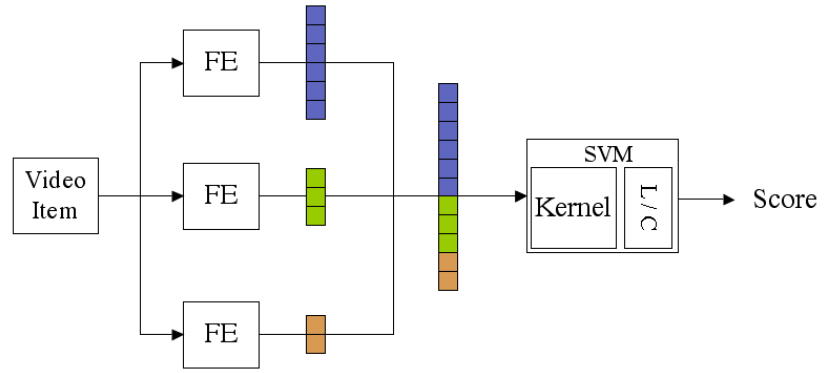
The most commonly used kernel function with SVM classifier in multimedia indexing is the RBF kernel defined as follow:

$$K(\mathbf{x}, \mathbf{y}) = e^{-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{2\sigma^2}}$$

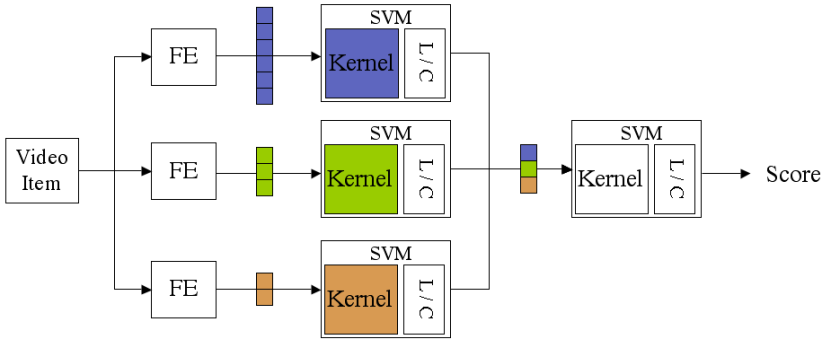
where $\| \, . \, \|$ denotes the L_2 norm, $\mathbf{x}$ and $\mathbf{y}$ are two sample vectors, and σ the width of the Gaussian kernel, generally determined using cross-validation. RBF kernels have exhibited good generalization properties in many classification problems. However, the use of a simple Euclidian distance implies small variations on the kernel value in high dimensional feature spaces.

3 Fusion schemes

We present in this section three fusion schemes inspired from the usual early and late fusion schemes. Those schemes use a classifier to learn the relations between modality components at different abstraction levels.



(a) Early Fusion



(b) Late Fusion

Fig. 1. Classical Early and Late Fusion schemes

Figure 1 describes the process of early and late fusion schemes. The feature extraction (FE) process extracts and creates a vector for each modality of the video item. We show the SVM process as two main steps: first, the construction of the Kernel, then the Learning or Classification (L / C) processes aims to assign a classification score to the video item.

Merging all the descriptors into a single flat classifier leads to a fully integrated fusion strategy since the fusion classifier obtains all the information from all sources. The advantage of such a scheme is its capacity to learn the regularities formed by the components independently from the modalities. Also, it is easy to use as it just consists in concatenating the various data in a single vector. The main disadvantage is the use of a unique approach (classifier and/or kernel) to merge different types of information. Assuming a RBF kernel and two sample vectors $\mathbf{x}$ and $\mathbf{y}$ from sets of features 1 and 2, the classical early fusion scheme leads to the following kernel:

$$K(\mathbf{x},\mathbf{y}) = e^{-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{2\sigma^2}} = e^{-\frac{\|\mathbf{x}_1-\mathbf{y}_1\|^2+\|\mathbf{x}_2-\mathbf{y}_2\|^2}{2\sigma^2}}$$
$$= e^{-\frac{\|\mathbf{x}_1-\mathbf{y}_1\|^2}{2\sigma^2}} e^{-\frac{\|\mathbf{x}_2-\mathbf{y}_2\|^2}{2\sigma^2}}$$

This formulation shows that using a SVM classifier with RBF kernels, an early fusion scheme is equivalent to multiply unimodal kernels which share the same σ parameter. The σ parameter is often fixed by cross validation, it is then optimal for the concatenated vectors, but not necessary for each modality

A late Fusion is performed on top of several classifiers. It has been presented using different formalisms, such as meta-classification which aims to re-classify the classification results made by other classifiers [9]. The closest theory to illustrate a late Fusion is the Stacking Ensemble learning [16] which is part of the ensemble methods [5]. The idea behind Ensemble learning methods (e.g. bagging, boosting, stacking) is to improve the generalization by training more than one model on each problem (e.g. train 10 SVM instead of just one) and then to combine their predictions by averaging, by voting or by other methods. Using staking, the combination is achieved by a final classifier which provides the final result. Hence, in the context of multimedia indexing, the late fusion scheme consists in performing a first classification separately on each modality and then in merging the outputs using a higher level classifier. In such a way, in contrast with the early fusion, one can use different classifier algorithms and different training sets according to the modalities. Furthermore, the late fusion scheme also allows to combine various classifiers for the same modality. However, the significant dimensional reduction induced by the stacked classifiers might be a disadvantage as the fusion classifier cannot fully benefit from the correlation among the sources of information.

3.1 Kernel Fusion

Kernel combination is a current active topic in the field of machine learning. It takes benefit of Kernel-based classifier algorithms. Advantages of merging modalities at kernel level are numerous. First, it allows to choose the kernel functions according to the modalities. For instance, histograms of colors can take advantage of specific histogram matching distances. Likewise, textual modality can be categorized using appropriate kernels such as String Kernels [10] or Word-Sequence kernels [4].

Kernel fusion also allows to model the data with more appropriate parameters. Merging modalities using an early fusion scheme leads to model the data using a single kernel function. Consequently, when using a RBF kernel, a single σ parameter is expected to "fit" properly the sample vectors relations, whereas it makes much more sense to train a combined RBF kernel using one σ per modality. Combination of unimodal kernels leads to keep as much information as possible from each modality. A combined RBF kernel has the following form:

$$K_c(\mathbf{x},\mathbf{y}) = F(K_m(\mathbf{x}_m,\mathbf{y}_m)_{(1\leq m\leq M)})$$

where $K_c(\mathbf{x}, \mathbf{y})$ is the combined kernel value for samples $\mathbf{x}$ and $\mathbf{y}$, $(K_m)_{1 \leq m \leq M}$ are the considered unimodal RBF kernels, F is the combining function over the M modalities, $\mathbf{x}_m$ and $\mathbf{y}_m$ are the sample vectors for modality m. Figure 2 shows the kernel fusion process, the unimodal kernels are merged using a fusion function in order to create the multimodal kernel. Then, learning and classification steps aim to assign a classification score to the video item.

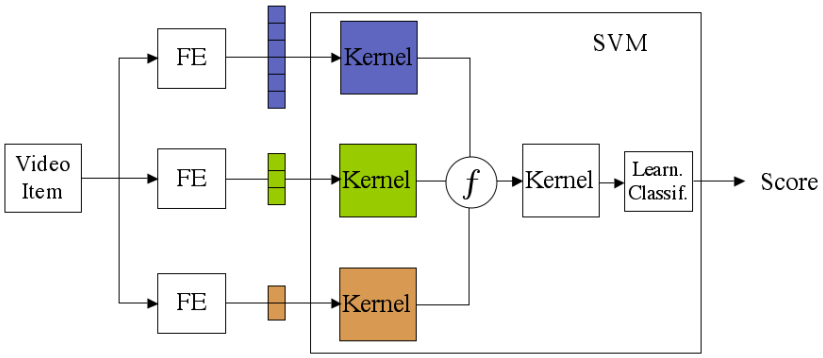


Fig. 2. Kernel Fusion scheme

One of the main issues in the current kernel research is the learning of such combined kernels. Called *Multiple Kernels Learning*, it aims to learn at the same time the parameters of all the unimodal kernels and the parameters of the combining function [15]. In our experiments, we used a very simple strategy to create combined kernels. The following algorithm describes the steps to simply create combined kernels:

1. Construct each unimodal kernels K_m,
2. Perform cross-validation on each unimodal kernels to fix their parameters,
3. Construct the combined kernel using the F combining function,
4. Perform cross-validation to optimize the parameters of F.

This algorithm assumes that the best parameters of unimodal kernels are suitable enough to allow efficient generalization of the combined kernel.

Combining individual kernels using a product operator is highly comparable to the classic early scheme where feature vectors are just concatenated. The difference is that by performing kernel fusion, each modality m is associated to its own kernel parameters (ie: σ_m. Furthermore, due to the product operator, this combination might lead to sparse kernels and provide poor generalization. We used the sum operator instead of the product operator to try to avoid too sparse kernel representations. Summing unimodal kernels should be more suitable for

concept detection when extracted features from a single modality are noisy and lead to incorrect detection.

We actually combine unimodal kernels by linear combination (weighted sum). Using RBF unimodal kernels, combined kernels are defined by the following formula:

$$K_c(\mathbf{x}, \mathbf{y}) = \sum_m w_m \, e^{-\frac{\|\mathbf{x}_m - \mathbf{y}_m\|^2}{2\sigma_m^2}}$$

where σ_m is the RBF parameter of kernel m and w_m is the weight of the associated modality. The w_m's can be fixed *a priori* or by cross-validation. In the conducted experiments, we optimized the w_m's on the training set.

3.2 Normalized Early Fusion

The number of extracted features depends upon the modalities and the type of the features. Hence, an early fusion scheme based on simple vector concatenation is much affected by the vector which has the highest number of inputs. Such fusion should have an impact on the classification, especially with a RBF kernel which is based on Euclidian distance between each training sample.

In traditional SVM implementation, a normalization process is integrated and aims to transform each input in the same range (e.g. [0..1], [−1..1]) in order to unbias the Euclidian distance. But, for the scope of merging features, this normalization doesn't take into account the number of input from the source features. The goal of normalized early fusion scheme is to avoid the problem of imbalanced features inputs by reprocessing each feature vectors before concatenation. We normalized each entry of the concatenated vector so that the average norm of each source vector is about the same. The normalization formula becomes:

$$x_i\prime = \frac{x_i - min_i}{(max_i - min_i) \times \sqrt{Card(x_i)}}$$

where x_i is an input of the feature vector x, min_i and max_i are respectively the minimum and maximum value of the i^{th} input among the training samples and $Card(x_i)$ is the number of dimensions of the source vector of x_i.

3.3 Contextual-Late Fusion

Usual late fusion scheme first classify each concept using individual modalities and then merge the scores in a second layer of classifier. Here, we generalize this scheme by considering more than a single concept. Contextual information has been widely exploited in multimedia indexing [14, 11]. Here, the second layer (stacked) classifier is able to exploit the contextual relation between the different concepts. This proposed scheme merges each unimodal classification score from a set of several concepts, in order to exploit both multimodal and conceptual contexts.

Assume that we have M modalities (e.g. visual, audio and text) and C concepts (e.g. Car, Face, Outdoor, Bus, etc). The stacked classifier merges M scores to classify the C concepts in the classic late fusion scheme. The late context fusion scheme merges $M \times C$ classification scores to classify the C concepts.

4 Experiments

We have evaluated and compared the presented fusion schemes in the framework of the TRECVID'06 evaluation campaign. The objective of the "high level feature extraction task" is to find video shots containing a visual appearance of 39 predefined concepts (high level features). For each concept, an ordered list of 2000 relevant shots should be returned by the competing systems. The Inferred Average Precision (IAP) [17] on the returned lists computed using the `trec_eval` tool is used as the evaluation metric. We compare the three proposed fusion schemes with the commonly used early and late fusion schemes, as well as with unimodal approaches. We have extracted features from visual and textual modalities; we present them in the following section.

4.1 Visual and text features

The features used in this evaluation are mid-level semantic features. Visual features are based on the concatenation of several intermediate concept classification scores detected at a patch level. Those visual "local concepts" are automatically extracted in each key frame, which are split into 260 (20×13) overlapping patches of 32×32 pixels. Local descriptors (low-level features) include: color (9 color momentum in RGB space), texture (8 orientation $\times$ 3 scales Gabor filters) and motion vector (extracted by optical flow). An SVM classifier is trained in order to detect a set of 15 visual concepts (eg: vegetation, sky, skin, etc.) selected from the LSCOM ontology. Those intermediate concepts have been selected as they can be extracted at patch level. For each of the 39 concepts, we manually associated a subset of 6 intermediate visual concepts. Thus, visual feature vectors contain 1560 dimensions (6×260).

Text features are derived from speech transcription result. We used 100 categories of the TREC Reuters collection to classify each speech segment. The advantages of extracting such concepts from the Reuters collection are that they cover a large panel of news topics like the TRECVID collection and they are obviously human understandable. Thus, they can be used for video search tasks. Examples of such topics are: economics, disasters, sports and weather. The Reuters collection contains about 800000 text news items in the years 1996 and 1997.

We constructed a vector representation for each speech segment by applying stop-list and stemming. Also, in order to avoid noisy classification, we reduced the number of input terms. While the whole collection contains more than 250000 terms, we have experimentally found that considering the top 2500 frequently

occurring terms gives the better classification results on Reuters collection. We built a prototype vector of each topic category on Reuters corpora and apply a Rocchio classification on each speech segment. Such granularity is expected to provide robustness in terms of covered concepts as each speaker turn should be related to a single topic. Our assumption is that the statistical distributions of the Reuters corpus and of TRECVID transcriptions are similar enough to obtain relevant results. Finally, the vector of text features has 100 dimensions. More explanation about those features can be found in [2].

4.2 Comparison of fusion schemes

The goal of the experiment is to study how imbalanced input features and large difference on the performance of unimodal classifiers are managed by the various fusion schemes. We show the results for unimodal runs and we compare all proposed fusion schemes and the usual early and late schemes. The 20 following concepts have been assessed for the TRECVID'06 evaluation campaign: Sports, Weather, Office, Meeting, Desert, Mountain, Waterscape, Corporate Leader, Police / Security, Military, Animal, Computer / TV Screen, US Flag, Airplane, Car, Truck, People Marching, Explosion / Fire, Maps, Chart.

The results presented in this paper are based on those 20 concepts. We do not study here each individual concepts result due to lack of space. The table 1 shows the Mean Inferred Average Precision (MIAP) obtained from the 20 assessed concepts. The two first entries refer to the unimodal runs, the two following correspond to the state of the art fusion schemes. The three fusion schemes described in this paper are shown in bold. We also show the median MIAP obtained from all of the TRECVID'06 participants.

Visual	0.0634
Text	0.0080
Classical Early Fusion	0.0735
Classical Late Fusion	0.0597
Normalized Early Fusion	**0.0884**
Kernel Fusion	**0.0805**
Contextual Late Fusion	**0.0753**
Median	0.0680

Table 1. Mean IAP of the 20 TRECVID'06 concepts

Unimodal runs :
We observe that the two unimodal runs are very different in terms of accuracy; the visual based classification is almost 7 times higher than text based concept detection. This is probably due to the nature of the assessed concepts, which seems to be hard to detect using text modality. The difficulty to detect concepts from the text modality is also probably due to the poor quality of automatic speech transcription (and translation) in some videos of the collection. This

point is actually interesting for the evaluation of the ability of the various fusion schemes to handle such heterogeneous data. The features we want to merge lead to different accuracies and are also imbalanced regarding the number of input features.

Classic Early and Late fusion schemes :
The two classical fusion schemes do not merge unimodal features similarly. While early fusion is able to outperform both unimodal runs, the late fusion scheme achieves poorer accuracy than the visual run. It might be due to the low number of dimensions handled by the stacked classifier. The early fusion scheme exploits context provided by all of the local visual features and the textual features. The gain obtained by such fusion means that those two modalities provide distinct kind of information. The merged features are, somehow, complementary.

Early based fusion schemes :
The gain obtained by the normalized fusion schemes is the most important compared to other fusion schemes. Processing the unimodal features by re-equilibrating them according to the number of dimensions is determining factor in order to significantly outperform unimodal runs. In such a way, despite the different number of dimensions, both the visual and textual modalities have the same impact on concept classification. This normalization process leads to a gain of almost 17% (in MIAP) comparing to the classic early fusion scheme, which simply normalizes input in a common range, and 28% comparing to the best unimodal run.

The gain obtained by the kernel fusion scheme is less significant than the gain obtained by the normalized fusion run. However, when comparing to the classic early fusion, it seems that a combination using sum operator leads to better accuracy than multiplying kernels (which is somehow what the classic early fusion do). Furthermore, it is important to notice that the σ parameters are selected first by cross-validation on unimodal kernels and that we optimize the linear combination separately. We can expect that an integrated framework which learns simultaneously σ_m and w_m parameters should lead to better results.

Contextual-Late fusion scheme :
Contextual-Late fusion is directly comparable with the classical late fusion scheme. This fusion scheme take into account the context from the score of other concepts detected in the same shot. By doing so, the context from other concepts leads to a gain of 26%. Furthermore, we observe that the MIAP obtained using the late contextual fusion scheme is almost the same as the one obtained for the classical early fusion scheme. In order to go further in this study, it could be interesting to evaluate the impact of the number and/or accuracy rate of concepts used in the context.

We notice that both of unimodal runs lead to poorer accuracy than the median of TRECVID'06 participants. This may be due to the basic and not so optimized features used in our experiments. However, the gain induced by the three fusion schemes presented in this paper lead to better accuracy than the median. We think that an optimization in the choice of descriptors for each modality could enhance the accuracy rate of both unimodal and multimodal runs.

5 Conclusion

We investigated three fusion schemes derived from the classical early and late fusion schemes when using SVM classifier. We have shown that all of the presented strategies perform in average better than the best unimodal run on the concept detection task of TRECVID'06. Furthermore, those fusion schemes outperform the median of TRECVID'06 participants over all of their runs. Kernel fusion schemes make it possible to take advantage of individual modalities, with a set of suitable parameters. Normalized early fusion is a good way to re-equilibrate the influence of individual modalities. Finally, the Contextual-Late fusion allows integration of context information from unimodal classification score of other concepts.

We studied influences of those fusion schemes on a set of 20 concepts, and did not analyzed individual concepts variations. As argued in [14], it is possible that one strategy performs differently than other depending the nature of the concepts. It could be interesting to go further in this direction. Also, the nature of the combined feature differs depending of the fusion schemes: early fusion is based on low- or intermediate-level features, where late fusion merges unimodal classification scores of high-level features. It could be interesting to merge those two heterogeneous kind of features in an integrated fusion scheme.

Bibliography

[1] S. Ayache, G. Quénot, J.Gensel, and S. Satoh. CLIPS-LSR-NII experiments at TRECVID 2005. *In proceedings of TRECVID Workshop*, 2005.

[2] S. Ayache, G. Quénot, J.Gensel, and S. Satoh. Using topic concepts for semantic video shots classification. *In proceedings of CIVR*, 2006.

[3] S. Boughorbel, J. Tarel, and N. Boujemaa. Conditionally positive definite kernels for SVM based image recognition. *In proceedings of ICME*, 2005.

[4] N. Cancedda, E. Gaussier, C. Goutte, and J.-M. Renders. Word-sequence kernels. *Journal of Machine Learning Research*, 2003.

[5] T. G. Dietterich. Ensemble methods in machine learning. In *Lecture Notes in Computer Science*, 2000.

[6] G. Iyengar and H. Nock. Discriminative model fusion for semantic concept detection and annotation in video. *In proceedings of ACM Multimedia*, 2003.

[7] W. Kraaij, P. Over, T. Ianeva, and A. F. Smeaton. *TRECVID 2006 – An Introduction*, 2006. http://www-nlpir.nist.gov/projects/tvpubs/tv6.papers/tv6intro.pdf.

[8] G. R. Lanckriet, M. Deng, N. Cristianini, M. I. Jordan, and W. S. Noble. Kernel-based data fusion and its application to protein function prediction in yeast. In *Proceedings of the Pacific Symposium on Biocomputing*, pages 300–311, 2004.

[9] W. Lin, R. Jin, and A. Hauptmann. Meta-classification of multimedia classifiers. In *Proceedings of First International Workshop on Knowledge Discovery*, 2002.

[10] H. Lodhi, J. Shawe-Taylor, N. Cristianini, and C. J. C. H. Watkins. Text classification using string kernels. *NIPS*, 2000.

[11] M. Naphade. On supervision and statistical learning for semantic multimedia analysis. *Journal of Visual Communication and Image Representation*, 2004.

[12] J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

[13] C. Snoek, M. Worring, and A. Smeulders. Early versus late fusion in semantic video analysis. *In proceedings of ACM Multimedia*, 2005.

[14] C. G. Snoek, M. Worring, J.-M. Geusebroek, D. C. Koelma, F. J. Seinstra, and A. W. Smeulders. The semantic pathfinder for generic news video indexing. In *Proceedings of ICME*, 2006.

[15] S. Sonnenburg, G. Ratsch, and C. Schafer. A general and efficient multiple kernel learning algorithm. *In proceedings of NIPS*, 2005.

[16] D. H. Wolpert. Stacked generalization. *Journal of Neural Networks*, 1990.

[17] E. Yilmaz and J. A. Aslam. Estimating average precision with incomplete and imperfect judgments. In *Proceedings of CIKM*, 2006.

[18] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: A comprehensive study. *Beyond Patches workshop, In proceedings of conjunction with CVPR*, 2006.